DIVERSITY AUGMENTED CONDITIONAL GENERATIVE ADVERSARIAL NETWORK FOR ENHANCED MULTIMODAL IMAGE-TO-IMAGE TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditional generative adversarial networks (cGANs) play an important role in multimodal image-to-image translation. We propose **Div**ersity **Aug**mented conditional **G**enerative **A**dversarial **N**etwork (DivAugGAN), a highly effective solution to further resolve the mode collapse problem and enhance the diversity for the generated images. DivAugGAN functions as a regularizer to maximize the distinction of the generating samples when different noise vectors are injected. We also exert extra constraint on the generator to ensure the relative variation consistency in the translation process. This guarantees that the changing scale of the generated images in the image space is coherent to the difference of the injected noise vectors in the latent space. It also reduces the chances to bring about unexpected mode override and mode fusion issues. Experimental results on both two-domain and multi-domain multimodal image-to-image translation tasks demonstrate its effectiveness. DivAugGAN leads to consistent diversity augmentations and visual quality improvements for the developed models. We also achieves state-of-the-art performances on multiple datasets in terms of widely used quantitative evaluation metrics. DivAugGAN can be easily integrated into any objectives in conditional generative models as a regularizer for diversity augmentations and quality enhancements without any additional computation overheads compromise. The source code and pre-trained models of our method will be available at `https://github.com/anomymous-gan/DivAugGAN`..

1 INTRODUCTION

Generative models, i.e., generative adversarial networks (GANs) Goodfellow et al. (2014); Goodfellow (2016), variational autoencoders (VAEs) Kingma & Welling (2014); Doersch (2016); van den Oord et al. (2017) and regressive models van den Oord et al. (2016a); Salimans et al. (2017); van den Oord et al. (2016b), have been widely implemented to capture complex high-dimensional data distribution. Its conditional variants, conditional generative models (CGMs) Mirza & Osindero (2014), take additional contexts to learn the mapping function from input to output distributions. Many conditional generation works are built up by CGMs. For example, the conditional variants of GANs (cGANs) are widely applied in many image generation, synthesis and translation tasks Isola et al. (2017); Zhu et al. (2017a), and video prediction, synthesis and translation tasks as well Mathieu et al. (2016); Tulyakov et al. (2018); Villegas et al. (2017); Clark et al. (2019); Wang et al. (2018a; 2019); Hsieh et al. (2018). It also plays a vital role on boosting the development of image-to-image translation, the aim of which is to learn the mapping between different visual domainsKamil & Shaikh (2019); Lin et al. (2018). There many computer vision and graphics problem can be formulated as image-to-image translation tasks, such as mapping grascale images to color images (colorization) Zhang et al. (2016); Larsson et al. (2016), mapping low-resolution images to the high-resolution images (super-resolution) Dong et al. (2015); Ledig et al. (2017); Zhang et al. (2019); Wang et al. (2018c) , mapping the corrupted images of missing region into the complete image (image inpainting) Iizuka et al. (2017); Yu et al. (2018), changing the attributes of a given image to another (attribute editing) He et al. (2019); Liu et al. (2019); Lu et al. (2018); Wu et al. (2019), synthesizing the photo-realistic image from the label or edge (photo-realistic image synthesis) Isola et al. (2017); Wang et al. (2018b); Ledig et al. (2017); Wang et al. (2018c) and transferring the styles of one domain to another (style transfer) Gatys et al. (2016); Johnson et al. (2016); Huang & Belongie (2017); Luan et al. (2017); Park & Lee (2019).

It is quite challenging to learn the mapping between different visual domains with superior visual quality, scalability, and diversity, as aligned training image pairs are very difficult or even impossible to collect Choi et al. (2020); Kim et al. (2017); Yi et al. (2017); Zhu et al. (2017a). Mapping is usually not deterministic but inherently multimodal. Designing and learning such models is quite complicated, especially in the case of a large number of attributes, domains, and styles existing.

To handle the scalability, previous studies proposed a unified framework to learn the multi-domains mappings between all available domains by using a single generator. Among them, StarGANChoi et al. (2018), AttGAN He et al. (2019), and RelGAN Wu et al. (2019) take a domain label as an extra input to transform the conditional input image to the target domain. As a fixed, predetermined label is given to the generator, it is still inevitably to produce the deterministic mapping result per each output domain in such frameworks. Lin *et al.* explore domain supervision to explicitly identify the domain of input conditional image by a pre-trained classification network to achieve multi-domain image-to-image translation Lin et al. (2019).

Many works on image-to-image translation have been developed to diversify the styles Zhu et al. (2017b); Lee et al. (2018); Huang et al. (2018); Na et al. (2019); Mao et al. (2019); Yang et al. (2019). A straightforward approach is to inject noise vectors (introduce style variations), usually randomly sampled the normal distribution, to the generator together with the input conditional images (maintain main contents). For example, in SYNTHIA $\rightleftharpoons$ Cityscape image-to-image translation task Huang et al. (2018), the street scenes and contents, e.g., positions and decorates of the constructions, buildings, trees and cars in the streets, are determined by the conditional input images, while the injected noise vectors support to diversify the lighting, shadow, and road textures, etc. Such solutions may not work well in cGANs based framework, as for the commonly appeared mode collapse problems. Generators are likely to only produce images from several major modes in the distribution while ignore other modes. Mode seeking GAN (MSGAN) Mao et al. (2019) and diversity-sensitive GAN (DSGAN) Yang et al. (2019) propose a similar regularization term to maximize the ratio of the distance between the generated samples with respect to the difference between the injected latent codes, which has enforced the generator to produce the distinct mages. However, MSGAN and DSGAN only partially resolve the mode collapse issue in cGANs and may also bring about unexpected mode fusion or mode override problems, as they do not consider the relative variation constraint in the translation process.

In this work, we propose **Div**ersity **Aug**mented **GAN** (DivAugGAN) to further enhance the multimodality of cGANs based framework and prevent the occurrence of mode fusion or mode override problems for the image-to-image translation task. We use three different latent vectors, which are constructed by a randomly sampled latent vector, and two relative offsets, to produce three output images. We propose to not only maximize the distance between each image pair with respect to the distance between the corresponding latent vectors pair, but also exert additional constraints on the generator to minimize the relative variation of the difference between the produced images and difference between the injected latent codes. Hence, the generators are encouraged to produce distinct images and maintain the scale of relative variations as well. It also enhances the generator to gain more chances to reach the minor modes and even nearby local maxima, and produce samples from different modes to match well with the real data distribution. Dissimilar generated samples from the nearby minor modes or local maxima that satisfied the relative variation coherence constraint, which may be ignored in otherwise, provide gradients to the discriminators. In comparison to MSGAN and DSGAN, our DivAugGAN can be readily embedded into all cGAN frameworks for image generation, synthesis and translation with enhanced diversity without any additional computational overheads.

We validate the effectiveness of the proposed DivAugGAN regularizer through an extensive application on multiple different cGANs based frameworks for both of two-domain and multi-domain multimodal image-to-image translation tasks. We achieve state-of-the-art performance on multiple datasets, e.g., *cat*$\rightleftharpoons$*dog*, *summer*$\rightleftharpoons$*winter*, *alps seasonal transfer*, *image weather condition*, *high-quality animal faces* (*AFHQ*), and *WikiArts* datasets, in terms of both of the qualitative and quantitative evaluation. We employ the following metrics: i) we employ *Fréchet Inception Distance* (*FID*) Heusel et al. (2017) as a metric for visual quality evaluation; ii) we employ *Learned Perceptual Image Patch Similarity* (*LPIPS*) Zhang et al. (2018) for diversity assessment; iii) we use *the Number of Statistically-Different Bins* (*NDB*) to determine the relative proportions of samples fallen into clusters that predetermined by the real data; and iv) we use *Jensen-Shannon Divergence* (*JSD*) distance

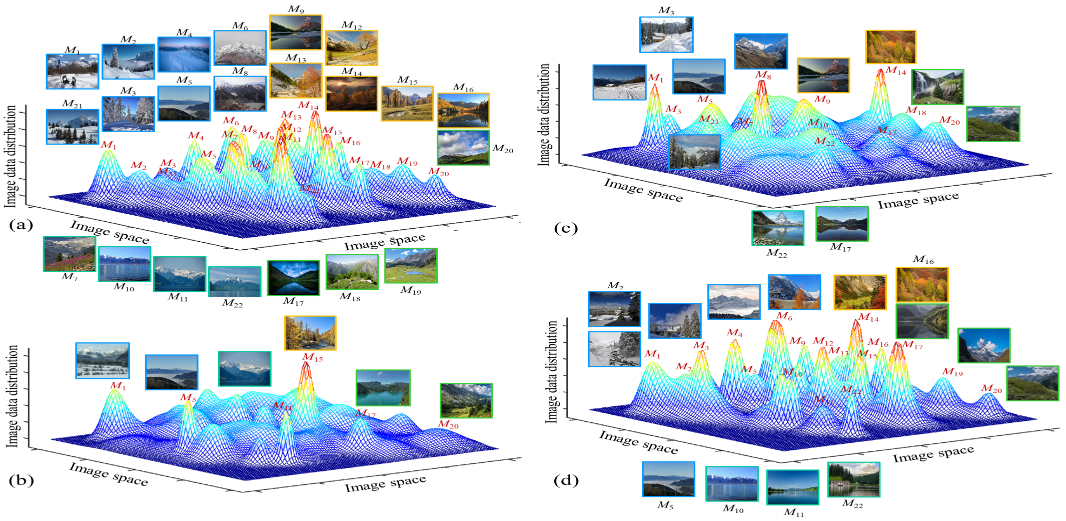

Figure 1: The main motivation: (a) Real data distribution contains numerous modes. (b) Generator is likely to focus on the major modes, while ignoring the minor modes, when mode collapse problem is appeared. Consequently, samples are produced from a few modes. (c) MSGAN and DSGAN Mao et al. (2019); Yang et al. (2019) simply maximize the distinction of the generating samples, when different noise vectors are injected, to alleviate the mode collapse problem. However, they fail to maintain relative variation, which may lead to mode override or mode fusion (e.g., nearby modes M_{14} and M_{16} are merged) issues. (d) Our DivAugGAN exert extra constraint on the generator to ensure the relative variation consistency in the translation process. This also guarantees that the changing sale of the generated images in the image space is coherent to the difference of injected noise vectors in the latent space. DivAugGAN is highly effectivesolution to further resolve the mode collapse problem and enhance the diversity.

Richardson & Weiss (2018) to measure the similarity between bin distributions. Experimental results show that the proposed DivAugGAN can function as a regularizer to facilitate the developed models to achieve the enhanced diversity without image quality and computational overhead compromise in both of the two-domain and multi-domain image-to-image translation.

In a nutshell, our main contributions in this work can be summarized as follows. 1) We propose DivAugGAN, which works as a regularizer to further suppress the mode collapse problem and reduce the possibility of bringing about unexpected mode fusion or mode override issues. It can be readily applied to enhance the diversity and improve the quality of the generated samples. 2) Extensive experiments demonstrate the universal effectiveness of DivAugGAN. We achieve state-of-the-art performance on multiple datasets with vastly different distributions in terms of qualitative and quantitative metrics, including, *FID*, *LPIPS*, *NDB*, and *JSD*. 3) The proposed DivAugGAN regularization scheme can be easily integrated into existing CGMs frameworks, with broad classes of the loss function, network architecture, and data modality. Extensive empirical results on image-to-image translation tasks demonstrate its effectiveness.

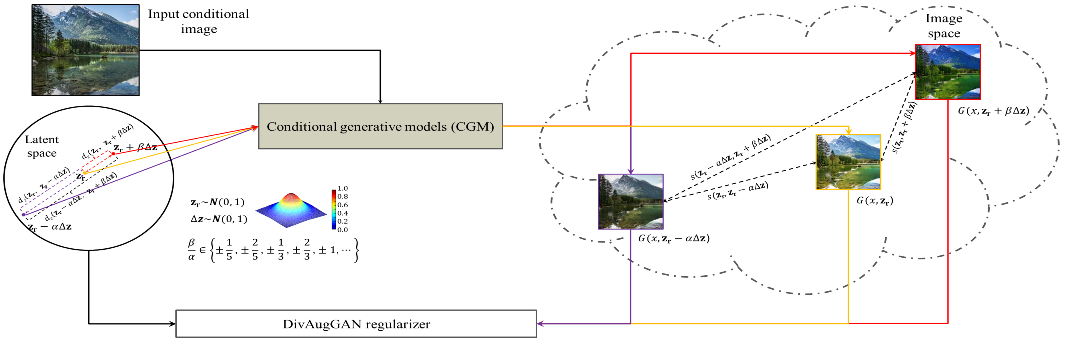
Input conditional image
Conditional generative models (CGM)
Latent space
$\mathbf{z_r} + \beta\Delta\mathbf{z}$
$\mathbf{z_r}$
$\mathbf{z_r} - \alpha\Delta\mathbf{z}$
$d_z(\mathbf{z_r}, \mathbf{z_r} + \beta\Delta\mathbf{z})$
$d_z(\mathbf{z_r}, \mathbf{z_r} - \alpha\Delta\mathbf{z})$
$d_z(\mathbf{z_r} - \alpha\Delta\mathbf{z}, \mathbf{z_r} + \beta\Delta\mathbf{z})$
$\mathbf{z_r} \sim \boldsymbol{N}(0,1)$
$\Delta\mathbf{z} \sim \boldsymbol{N}(0,1)$
1.0
0.8
0.6
0.4
0.2
0.0
$\frac{\beta}{\alpha} \in \left\{\pm\frac{1}{5}, \pm\frac{2}{5}, \pm\frac{1}{3}, \pm\frac{2}{3}, \pm 1, \cdots\right\}$
Image space
$G(x, \mathbf{z_r} + \beta\Delta\mathbf{z})$
$G(x, \mathbf{z_r})$
$G(x, \mathbf{z_r} - \alpha\Delta\mathbf{z})$
$s(\mathbf{z_r} - \alpha\Delta\mathbf{z}, \mathbf{z_r} + \beta\Delta\mathbf{z})$
$s(\mathbf{z_r}, \mathbf{z_r} - \alpha\Delta\mathbf{z})$
$s(\mathbf{z_r}, \mathbf{z_r} + \beta\Delta\mathbf{z})$
DivAugGAN regularizer

Figure 2: DivAugGAN functions as a regularizer in the CGMs framework. DivAugGAN encourages the generator to explore the unseen distant image space. It also enforces the scale of the relative variation for the generated images, which is consistent with the difference between the injected latent codes. As a result, the discriminator not only pays attention to the generated images from the minor modes, but also gain the possibility to differentiate little variation between the nearby local maxima. This reduces the chances of bringing about mode override and mode fusion problems.

2 Diversity augmented conditional image synthesis

2.1 Preliminaries

In the task of image-to-image translation, cGANs are applied on learning the conditional mapping function G to generate the output image $y \in \mathcal{Y}$ conditioned on an input image $x \in \mathcal{X}$, where $\mathcal{X}$ and $\mathcal{Y}$ represent the input and output image space, respectively Yang et al. (2019). In multimodal image synthesis and translation tasks, an input image x is mapped to multiple distinct outputs with different encoded latent codes $z \in \mathcal{Z}$ Zhu et al. (2017b). cGANs learn such multimodal mapping by alternatively updating the generator G and discriminator D to solve the following mini-max problem Goodfellow et al. (2014); Goodfellow (2016); Isola et al. (2017):

$$\min_G \max_D \mathcal{L}_{cGANs}(G, D) = \mathbb{E}_{x,y}\left[\log D\left(x, y\right)\right] + \mathbb{E}_{x,z}\left[\log\left(1 - D(x, G(x, z))\right)\right]. \tag{1}$$

Theoretically, through adversarial training, the gradients from the discriminator D progressively guide the generator G to produce samples with the distribution similar to the real data. However, in practice, the major modes are much more likely to be favored than the minor modes in the training process Mao et al. (2019), due to irregular data distribution in the mapped space. As a result, it may be very difficult or even impossible to successfully generating samples from the ignored minor modes, illustrated in 1 (b), which also leads to the well-known mode collapse problem Salimans et al. (2016); Srivastava et al. (2017). Extensive studies have been presented to resolve such issue in both of standard and conditional GANs, such as incorporating the mini-batch statistics into the discriminator Salimans et al. (2016), employing the improved divergence metrics, objective functions, and optimization processes to smooth the loss of the discriminator Arjovsky et al. (2017); Gulrajani et al. (2017); Mao et al. (2017); Odena et al. (2018); Heusel et al. (2017); Srivastava et al. (2017); Miyato et al. (2018), and introducing auxiliary networks, as multiple generators or discriminators with weight-sharing mechanism Liu & Tuzel (2016); Ghosh et al. (2018); Hoang et al. (2018); Nguyen et al. (2017); Che et al. (2016), extra encoders Dumoulin et al. (2017); Donahue et al. (2017); Larsen et al. (2016) and additional classifier Odena et al. (2017); Lin et al. (2019), etc.

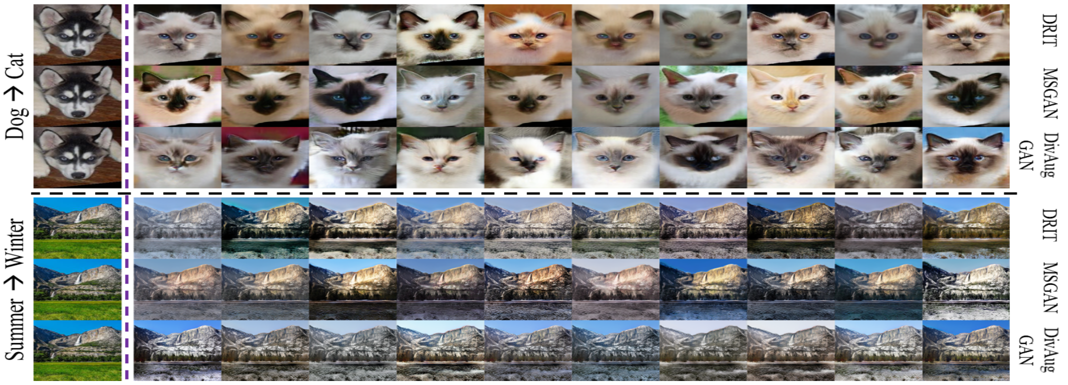

Figure 3: Qualitative comparisons of DivAugGAN with DRIT and MSGAN on *dog* → *cat*, and *summer* → *winter* for two-domain multimodal image-to-image translation tasks. DivAugGAN generates images with much more variation (diverse color, shape, light, and orientation) over DRIT and MSGAN. Complete results in the supplementary.

In the image-to-image translation tasks, the generator may be prone to focus on the high-dimensional structured conditional context, while ignore the low-dimensional stochastic latent codes $\mathbf{z}$ in some extreme cases, and learn a deterministic mapping from x to y. When encounter the mode collapse problem, the generator may map two different $\mathbf{z_s}, \mathbf{z_t}$ latent codes into the same mode, as shown in Fig.1 (b). Hybrid model of cGAN and VAE with random injected latent codes is the first work to address mode collapse issue in cGANs based framework for multimodal image-to-image translation. Specifically, Zhu *et al.* design an invertible generator in BiCycleGAN with an additional encoder network for latent code reconstruction from the generated image Zhu et al. (2017b). Domain-specific decoders are developed to interpret the latent codes for generating images with various styles in multimodal image translation by Lee *et al.* Lee et al. (2018) and Huang *et al.* Huang et al. (2018), respectively. Odena *et al.* propose a regularization method to clamp the generator Jacobian within a certain range Odena et al. (2018). Sharing a similar idea as Odena et al. (2018), Yang *et al.* presented DSGAN with an objective function to simply maximize the norm of the generator gradient with an optional upper-bound Yang et al. (2019). Mao *et al.* proposed MSGAN with an additional mode seeking regularization term on the generator to maximize the ratio of the distance between the produced images with respect to the distance between the injected latent vectors. All such regularization methods only encourage the generator to explore the distant space to gain more chances to hit the far-away minor modes, which enforces the generator to produce distinct outputs, illustrated in Fig.1 (c). However, they fail to maintain the relative variation consistency. In other words, the distance between the distinct generated images in the image space is not consistent with the difference between the injected latent codes in the latent space. Such incoherence may lead to mode override or mode fusion issues, as shown in Fig.1 (c).

2.2 Diversity augmented conditional generative adversarial networks

In this work, we propose DivAugGAN to further resolve the mode collapse issue, and keep away from bringing about the unexpected mode override or mode fusion problems. We not only utilize stochastic features of the injected latent codes to produce distinct samples, but also employ extra constraint to maintain the relative variation of the generated images that is consistent with the difference between the injected latent codes. Figure 2 presents the framework. Given the input conditional image x, let latent vector, $\mathbf{z_r}$, from the latent space $\mathcal{Z}$ to generate image, $G(x, \mathbf{z_r})$ on learning the mapping from conditional input $\mathcal{X} \times \mathcal{Z}$ to the output image space $\mathcal{Y}$. By given another reference latent code, i.e., $\alpha\Delta\mathbf{z}$, to reflect the change scale, we can define $d_z\{\mathbf{z_r}, \mathbf{z_r} - \alpha\boldsymbol{\Delta}\mathbf{z}\}$ as the distance between two latent codes $\mathbf{z_r}$ and $\mathbf{z_r} - \alpha\boldsymbol{\Delta}\mathbf{z}$, and $d_I\{G\left(x, \mathbf{z_r}\right), G\left(x, \mathbf{z_r} - \alpha\boldsymbol{\Delta}\mathbf{z}\right)\}$ as the distance between two produced images, i.e., $G(x, \mathbf{z_s})$ and $G(x, \mathbf{z_r} - \alpha\boldsymbol{\Delta}\mathbf{z})$, respectively. Here α is a scale factor and $\Delta\mathbf{z}$ is the reference quantity. For simplicity, we use $\delta(\mathbf{z_r}, \mathbf{z_r} - \alpha\boldsymbol{\Delta}\mathbf{z})$ and $s(\mathbf{z_r}, \mathbf{z_r} - \alpha\boldsymbol{\Delta}\mathbf{z})$ to represent the distance ratio $\frac{d_I\{G(x,\mathbf{z_r}),G(x,\mathbf{z_r}-\alpha\boldsymbol{\Delta}\mathbf{z})\}}{d_\mathbf{z}\{\mathbf{z_r},\mathbf{z_r}-\alpha\boldsymbol{\Delta}\mathbf{z}\}}$, and image distance $d_I\{G\left(x, \mathbf{z_r}\right), G\left(x, \mathbf{z_r} - \alpha\boldsymbol{\Delta}\mathbf{z}\right)\}$. Note that relative variation coherence is ruled out in Mao et al. (2019); Yang et al. (2019), as their generators are trained with a regularization term to simply maximize the distance ratio or the image distance. Evidently, it encourages the generator to explore some distant major modes by with compromises of identifying nearby local maxima and minor modes. To resolve their limitations, we propose DivAugGAN regularization scheme on the generator to maximize the distinction of the generated samples with different injected noise vectors and ensure the relative variation consistency of the image space and latent space as well:

$$\begin{aligned}\mathcal{L}_{da} = \max_G \mathbb{E}_{\mathbf{z_r},\boldsymbol{\Delta}\mathbf{z}}\{&\lambda_1[\delta(\mathbf{z_r}, \mathbf{z_r} - \alpha\boldsymbol{\Delta}\mathbf{z}) + \delta(\mathbf{z_r} + \beta\boldsymbol{\Delta}\mathbf{z}, \mathbf{z_r}) + \delta(\mathbf{z_r} + \beta\boldsymbol{\Delta}\mathbf{z}, \mathbf{z_r} - \alpha\boldsymbol{\Delta}\mathbf{z})] \\ &- \lambda_2[s(\mathbf{z_r} + \beta\boldsymbol{\Delta}\mathbf{z}, \mathbf{z_r}, \mathbf{z_r} - \beta\boldsymbol{\Delta}\mathbf{z}) + s(\mathbf{z_r} + 2\beta\boldsymbol{\Delta}\mathbf{z}, \mathbf{z_r} + \beta\boldsymbol{\Delta}\mathbf{z}, \mathbf{z_r}) \\ &+ s(\mathbf{z_r}, \mathbf{z_r} - \alpha\boldsymbol{\Delta}\mathbf{z}, \mathbf{z_r} - 2\alpha\boldsymbol{\Delta}\mathbf{z})]\}.\end{aligned} \tag{2}$$

where α and β are two scale factors, $\mathbf{z_r}$, and $\Delta\mathbf{z}$ are two latent codes randomly sampled from normal distribution $\mathcal{N}\left(0, 1\right)$, function as a reference and control relative change, respectively, $\| \cdot \|_1$ represents the L_1 norm. The first three terms in Eq.(2) encourage the generator to explore the unseen far-away image space to elevate chance of hitting the distant modes; the latter three terms in Eq.(2) enforce additional constraints on the generated images to ensure the changing scale in the image space is coherent to the variation of the injected latent codes in the latent space. The discriminator gain possibility to differentiate tiny difference of the nearby local maxima and increase the opportunities on using minor modes to generate images. It also minimizes the chance to bring about mode override or mode fusion problems, illustrated in Fig.1 (d).

2.3 Analysis of the diversity augmented regularization

DivAugGAN introduces a novel regularization for cGANs to promote local sensitivity. It directly augments the diversity of generated samples, i.e., $G(\mathbf{x}, \mathbf{z_r})$ with the latent style code $\mathbf{z_r}$. A large norm of the first-order derivative ensures the sensitive responses to style codes, and a moderate norm of the second-order derivative encourages weak decay of the sensitivity. Motivated from this point, we formulate the DivAugGAN regularizer as:

$$\mathcal{L}_{da} = \max_G \mathbb{E}_{\mathbf{z_r}} \left\{ \lambda_1 \left\| \frac{\partial G(\mathbf{x}, \mathbf{z_r})}{\partial \mathbf{z}} \right\| - \lambda_2 \left\| \frac{\partial^2 G(\mathbf{x}, \mathbf{z_r})}{\partial \mathbf{z}^2} \right\| \right\}. \tag{3}$$

We employ *finite difference methods* to approximate above norms with the average norms of the corresponding *directional derivatives* from the data pairs $(G(\mathbf{x}, \mathbf{z_r}), \mathbf{x}, \mathbf{z_r}, \boldsymbol{\Delta}\mathbf{z})$ along any random direction $\mathbf{v} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_k)$. For a multivariate function $g(\mathbf{x}) : \mathbb{R}^\ell \to \mathbb{R}$ and a directional vector $\mathbf{v} \in \mathbb{R}^\ell$, its first-order and second-order directional derivatives are closely related to the corresponding derivatives, i.e., $\|\mathbf{v}\|_2 \frac{\mathrm{d}g(\mathbf{x})}{\mathrm{d}\mathbf{v}} = \mathbf{v}^T \frac{\mathrm{d}g(\mathbf{x})}{\mathrm{d}\mathbf{x}}$, $\|\mathbf{v}\|_2^2 \frac{\mathrm{d}^2 g(\mathbf{x})}{\mathrm{d}\mathbf{v}^2} = \mathbf{v}^T \frac{\mathrm{d}^2 g(\mathbf{x})}{\mathrm{d}\mathbf{x}^2} \mathbf{v}$. Hence, we have the transformed regularization as:

Table 1: Quantitative comparisons of DRIT, MSGAN, and DivAugGAN in the *Cat* ⇌ *Dog* and Yosemite *Summer* ⇌ *Winter* datasets.

	Cat → Dog			Dog → Cat		
	DRIT	MSGAN	DivAugGAN	DRIT	MSGAN	DivAugGAN
FID ↓	33.17±0.63	17.76±0.15	**16.36±0.53**	22.98±0.34	25.72±0.82	**20.91±0.37**
LPIPS ↑	0.2142±0.0017	0.5080±0.0014	**0.5485±0.0014**	0.3984±0.0023	0.4330±0.0028	**0.4444±0.0010**
NDB ↓	31.67±1.33	**21.67±4.67**	30.00±1.00	21.00±1.00	**20.67±1.67**	25.33±1.67
JSD ↓	0.138±0.007	**0.078±0.001**	0.108±0.008	0.081±0.006	**0.074±0.006**	0.105±0.008
	Summer → Winter			Winter → Summer		
	DRIT	MSGAN	DivAugGAN	DRIT	MSGAN	DivAugGAN
FID ↓	47.85±0.23	47.77±0.05	**46.56±0.12**	42.97±0.15	40.58±0.14	**40.27±0.15**
LPIPS ↑	0.2216±0.0039	0.2756±0.0035	**0.2800±0.0002**	0.1817±0.0011	0.2257±0.0011	**0.2358±0.0011**
NDB ↓	26.33±1.67	23.33±2.33	**23.00±2.00**	22.33±1.67	20.33±0.67	**18.67±0.33**
JSD ↓	0.052±0.003	0.046±0.001	**0.041±0.003**	0.052±0.002	0.038±0.001	**0.038±0.001**

$$\mathcal{L}_{da} = \max_{G} \mathbb{E}_{\mathbf{z_r}} \left\{ \lambda_1 \mathbb{E}_{\mathbf{v}} \left[\left\| \frac{\partial G(\mathbf{x}, \mathbf{z_r})}{\partial \mathbf{v}} \right\| \right] - \lambda_2 \mathbb{E}_{\mathbf{v}} \left[\left\| \frac{\partial^2 G(\mathbf{x}, \mathbf{z_r})}{\partial \mathbf{v}^2} \right\| \right] \right\}, \tag{4}$$

where $\frac{\partial G(\mathbf{x},\mathbf{z_r})}{\partial \mathbf{v}}$ and $\frac{\partial^2 G(\mathbf{x},\mathbf{z_r})}{\partial \mathbf{v}^2}$ refer to the first-order and the second-order partial directional derivative of $G(\mathbf{x}, \mathbf{z})$ to $\mathbf{z}$, respectively. When ℓ_1 or ℓ_2 norm is employed, we can prove that the following proportional expression holds between the first-order norms[1]:

$$\mathbb{E}_{\mathbf{v} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_k)} \left[\left\| \frac{\partial G(\mathbf{x}, \mathbf{z})}{\partial \mathbf{v}} \right\| \right] \propto \left\| \frac{\partial G(\mathbf{x}, \mathbf{z})}{\partial \mathbf{z}} \right\|. \tag{5}$$

It implies that average norms of directional derivatives can function as a surrogate for norms of derivatives, especially in formulating the regularization losses.

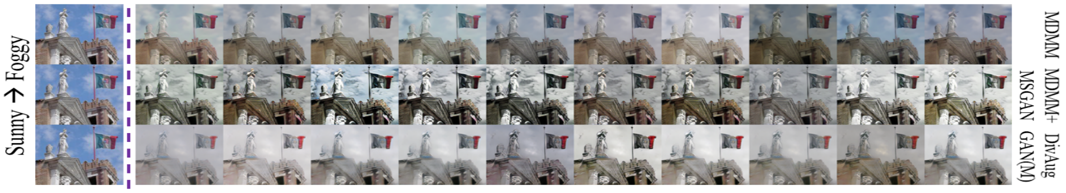

Figure 4: Qualitative comparisons of DivAugGAN (M) with MDMM and MSGAN integratd MDMM on *image weather condition* dataset for multi-domain translation. Our model generate images with enhanced diversity and superior visual quality. Only *sunny* → *foggy* translation result is shown; complete results in the supplementary.

3 Experiments

In this section, we demonstrate the effectiveness of the proposed DivAugGAN through extensive quantitative and qualitative evaluation on a variety of conditional image-to-image translation tasks. Both the two-domain and multi-domain are included. We apply the proposed DivAugGAN regularization objectives in the baseline models directly. Note that we simply modify the original objective functions to the proposed DivAugGAN regularization objectives, without changing the network architectures and hyper-parameters for fair comparison. All experiments are conducted using unseen images during the training phase.

3.1 Quantitative and Qualitative Results

To demonstrate the scalability and universal effectiveness of the proposed DivAugGAN, we evaluate it on a variety of datasets, listed as the following, i) Yosemite *Summer⇌Winter dataset* Zhu et al. (2017a), ii) *Dog⇌Cat* Lee et al. (2018), iii) *Alps seasonal transfer* Anoosheh et al. (2018), iv) *image weather conditions* Chu et al. (2017), v) *AFHQ* Choi et al. (2020) and vi) *WikiArts* Zhu et al. (2017a), with the widely used evaluation metrics, such as *FID*, *LPIPS*, *NDB*, and *JSD*, and compare it with the reference models for both of two-domain and multi-domain multimodal image-to-image translation tasks, including DRIT Lee et al. (2018), MSGAN Mao et al. (2019), MDMM Lee et al. (2020), and StarGANv2 Choi et al. (2020).

Two-domain multimodal image-to-image translation. We employ the network architectures of generator and discriminator from DRIT and MSGAN without modifications. We conduct experiments on i) *dog* ⇌ *cat*, and ii) Yosemite *summer* ⇌ *winter* datasets using the same hyperparameter settings

[1] The proof detail are presented in Appendix A.

Table 2: Quantitative comparisons of MDMM, MDMM with MSGAN regularizer, StarGANv2, MDMM with DivAugGAN regularizer, and StarGANv2 with DivAugGAN regularizer with *alps seasonal transfer*, *image weather conditions*, *AFHQ*, and *WiKiArts* datasets. *DSGAN regularizer is already intergrated into the StarGANv2 framework for diversification enhancement.

		MDMM	MDMM+MSGAN	DivAugGAN(M)	StarGANv2*	DivAugGAN(S)
Alps seasonal transfer	*FID* ↓	76.65±0.09	70.36±0.24	**66.83±0.02**	50.60±0.01	**44.09±0.16**
	LPIPS ↑	0.0863±0.0002	0.1397±0.0012	**0.1731±0.0017**	0.3043±0.0020	**0.4233±0.0005**
	NDB ↓	20.89±0.11	19.58±0.67	**18.11±0.39**	15.42±0.66	**13.97±0.72**
	JSD ↓	0.074±0.001	0.063±0.001	**0.059±0.001**	0.047±0.002	**0.041±0.001**
Image weather conditions	*FID* ↓	154.69±0.27	151.55±0.12	**142.06±0.24**	104.06±0.48	**97.23±0.31**
	LPIPS ↑	0.0669±0.0014	0.1109±0.0004	**0.1658±0.0016**	0.3962±0.0016	**0.4668±0.0005**
	NDB ↓	21.44±0.28	21.14±0.56	**19.94±0.52**	14.33±0.67	**11.58±0.16**
	JSD ↓	0.172±0.001	0.179±0.004	**0.166±0.003**	0.097±0.001	**0.081±0.001**
AFHQ	*FID* ↓	49.69±0.17	**21.34±0.09**	29.82±0.21	19.45±0.11	**18.69±0.10**
	LPIPS ↑	0.3098±0.0012	0.4520±0.0003	**0.4630±0.0005**	0.5007±0.0001	**0.5102±0.0005**
	NDB ↓	37.17±0.67	40.17±0.67	**38.50±0.17**	**29.50±0.50**	39.39±0.73
	JSD ↓	0.076±0.001	0.114±0.001	**0.111±0.001**	**0.051±0.001**	0.119±0.002
WikiArts	*FID* ↓	163.49±0.27	155.68±0.05	**116.25±0.08**	133.10±0.22	**101.52±0.10**
	LPIPS ↑	0.1256±0.0006	0.1796±0.0001	**0.6165±0.0007**	0.5261±0.0008	**0.6840±0.0006**
	NDB ↓	31.88±0.08	**32.42±0.17**	35.97±0.38	29.28±0.12	**27.13±0.18**
	JSD ↓	0.166±0.001	**0.123±0.002**	0.184±0.002	0.085±0.001	**0.081±0.001**

Table 3: Ablation study to investigate the effects of the proposed DivAugGAN regularizer for two-domain and multi-domain multimodal image-to-image translation.

		Cat → Dog{DivAugGAN}				Dog → Cat{DivAugGAN}			
DR	**RVC**	*FID* ↓	*LPIPS* ↑	*NDB* ↓	*JSD* ↓	*FID* ↓	*LPIPS* ↑	*NDB* ↓	*JSD* ↓
✗	✗	33.17±0.63	0.2142±0.0017	31.67±1.33	0.138±0.007	22.98±0.34	0.3984±0.0023	21.00±1.00	0.081±0.006
✓	✗	17.76±0.15	0.5080±0.0014	**21.67±4.67**	**0.078±0.001**	25.72±0.82	0.4330±0.0028	**20.67±1.67**	**0.074±0.0006**
✗	✓	26.67±0.33	0.4256±0.0016	31.00±1.00	0.113±0.006	22.67±0.67	0.4233±0.0019	21.00±1.00	0.114±0.0007
✓	✓	**16.36±0.53**	**0.5485±0.0014**	30.00±1.00	0.108±0.008	**20.91±0.37**	**0.4444±0.0010**	25.33±1.67	0.105±0.008
		Alps seasonal transfer {DivAugGAN(M)}				Image weather condition {DivAugGAN(S)}			
✗	✗	76.65±0.09	0.0863±0.0002	20.89±0.11	0.074±0.001	120.59±0.52	0.2432±0.0021	19.37±0.82	0.121±0.001
✓	✗	70.36±0.24	0.1397±0.0012	19.58±0.67	0.063±0.001	104.06±0.48	0.3962±0.0016	14.33±0.67	0.097±0.001
✗	✓	72.43±0.14	0.1235±0.0013	19.97±0.53	0.067±0.001	114.67±0.51	0.3138±0.0019	15.71±0.71	0.109±0.001
✓	✓	**66.83±0.02**	**0.1731±0.0017**	**18.11±0.39**	**0.059±0.001**	**97.23±0.31**	**0.4668±0.0005**	**15.16±0.16**	**0.081±0.16**

as such two baseline modals for a fair comparison. Table 1 summarizes the quantitative experimental results on Yosemite *dog* ⇌ *cat* and *summer* ⇌ *winter* datasets, respectively. DivAugGAN achieves consistent improvements on the diversity metric (higher *LPIPS* score) over DRIT and MSGAN. Qualitative comparison results in Figure 3 also confirm that DivAugGAN generates samples with superior diversity and visual quality over DRIT and MSGAN. Lower *FID* also indicates DivAugGAN consistently guides the distributions of the generated samples to match the real data.

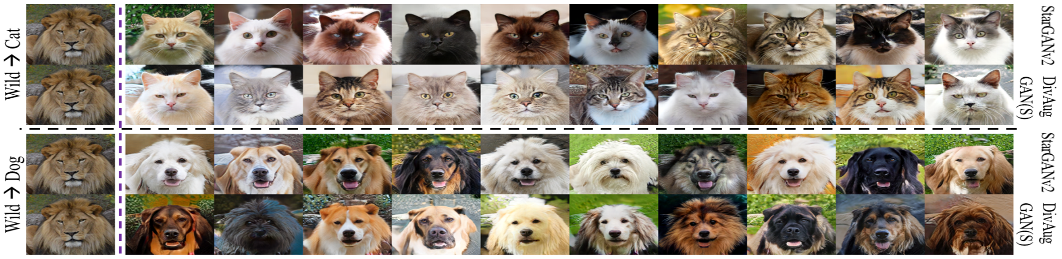

Figure 5: Qualitative comparisons of DivAugGAN (S) with StarGANv2 on *AFHQ* dataset for multi-domain multimodal translation. Our model can generate images with superior diversity. Only *wild* → *cat*, and *wild* → *dog* translation results are shown; complete results in the supplementary.

Multi-domain image-to-image translation. We employ the network architectures of generator and discriminator from MDMM Lee et al. (2020) and StarGANv2 Choi et al. (2020) without modifications. We conduct experiments on i) *alps seasonal transfer*, ii) *image weather conditions*, iii) *AFHQ*, and iv) *WikiArts* datasets using the same hyperparameter settings as such baseline models for a fair comparison. Note that the DSGAN regularizer has already been embedded into StarGANv2 for diversity enhancement. We integrate MSGAN regularization method into MDMM to obtain another model, named MDMM+MSGAN, for performance comparison. We build our DivAugGAN(M) and DivAugGAN(S) models, by simply replacing the MSGAN regularizer and DSGAN regularizer to our DivAugGAN regularizer in the MDMM+MSGAN, and StarGANv2 architectures, respectively. As quantitative experimental results exhibited in Table 2, both of the proposed DivAugGAN(M) and DivAugGAN(S) perform favorably against MDMM, MDMM+MSGAN, and StarGANv2 in almost all quantitative evaluation metrics in all of the four tasks. They consistently achieve higher *LPIPS*, lower *FID*, and *JSD* scores. Qualitative comparison results in Fig.4 and Fig.5 on *image weather condition*, and *AFHQ* datasets also demonstrate that the performance of DivAugGAN(M) is superior to MDMM and MDMM+MSGAN, and DivAugGAN(S) is superior to StarGANv2 as well. This is because the output samples from both of DivAugGAN(M) DivAugGAN(S) present much more diverse features and satisfying visual quality.

3.2 Ablation study

We run an ablation study to investigate the effects of each component in the proposed DivAugGAN regularizer in Eq. 2 for multimodal image-to-image translation. By progressively add DR terms (first three terms in Eq. 2), and RVC terms (latter three terms in Eq. 2) to the baseline architecture of DivAugGAN (DRIT for two-domain task and MDMM/StarGANv2 for multi-domain task), we verify that our superior results are benefited from both of them. The ablation study details on the *Dog⇌Cat* Lee et al. (2018), *Alps seasonal transfer* Anoosheh et al. (2018) and *image weather condition* Chu et al. (2017) datasets for two-domain and multi-domain multimodal image-to-image translation are presented in Table 3. When we add DR terms, the key quantitative metrics, i.e., *FID* (Cat→Dog: 33.17 → 17.76; Alps:76.65 → 70.36; weather:120.59 → 104.06), *LPIPS* (Cat→Dog: 0.2142 → 0.5080; Alps: 0.0863 → 0.1397; weather: 0.2432 → 0.3962), of the generated images are consistently improved. Similarly, the visual quality and diversity of the generated images are also benefited from adding RVC terms, i.e., *FID* (Cat→Dog: 33.17 → 26.67; Alps: 76.65 → 72.43; weather: 120.59 → 14.67), *LPIPS* (0.2142 → 0.4256, Alps:0.0863 → 0.1235; weather:0.2432 → 0.3138). When both of the DR terms and RVC terms are employed together, the quantitative results are further improved and the lowest *FID* (Cat→Dog/Dog→Cat: 16.36/20.91; Alps:66.83; weather:97.23) and highest *LPIPS* (Cat→Dog/Dog→Cat: 0.5485/0.4444; Alps:0.5485; weather:0.4668) are achieved. Such results justify that both of DR terms and RVC terms contribute to enhance the visual quality and diversity in the proposed DivAugGAN regularizer.

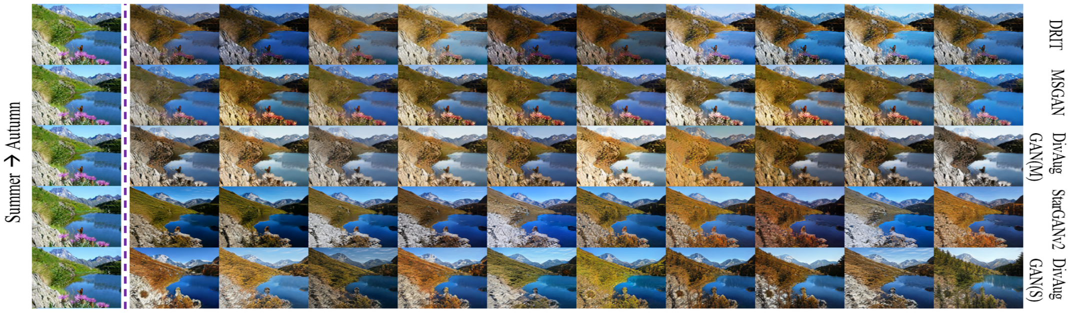

Figure 6: Qualitative comparison results on *alps seasonal transfer* dataset for multi-domain translation. Our model DivAugGAN(M) and DivAugGAN(S) consistently produce more diverse images over DRIT, MSGAN, and StarGANv2 with superior visual quality. Only *summer* → *autumn* translation result is shown; complete results in the supplementary.

4 Conclusions

We present DivAugGAN to further resolve the mode collapse problem. We exert diversity augmented regularization term on the generator to maximize the distinction of the producing samples and maintain the relative variation consistency in the translation process as well. This also helps to suppress modes override and mode fusion issues. The proposed regularization is simple, general, and can be readily integrated into the existing cGANs based framework without any additional computation overhead or network structures modification cost. The proposed regularization method has achieved state-of-the-art performance on multiple datasets with different distribution for both two-domain and multi-domain multimodal image-to-image translation tasks. Quantitative and qualitative experimental results also demonstrate the universal effectiveness of the proposed DivAugGAN, which is superior to previous MSGAN and DSGAN. Additionally, DivAugGAN can generate much more diverse images with higher visual quality. Appropriately selecting the distance metric for the produced images in the image space and carefully learning the ratio between α and β may further enhance the realism and diversity. It would be an interesting future work to explore the proper metrics to measure the mutual difference quantitatively between the generated images in image space.

References

Asha Anoosheh, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Combogan: Unrestrained scalability for image domain translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018. 6, 8

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. 4

Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li. Mode regularized generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2016. 4

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 6, 7

Wei-Ta Chu, Xiang-You Zheng, and Ding-Shiuan Ding. Camera as weather sensor: Estimating weather information from single images. *Journal of Visual Communication and Image Representation*, 2017. 6, 8

Aidan Clark, Jeff Donahue, and Karen Simonyan. Efficient video generation on complex datasets. *arXiv preprint arXiv:1907.06571*, 2019. 1

Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016. 1

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *International Conference on Learning Representations (ICLR)*, 2017. 4

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 2015. 1

Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. In *International Conference on Learning Representations (ICLR)*, 2017. 4

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

Arnab Ghosh, Viveka Kulharia, Vinay P. Namboodiri, Philip H.S. Torr, and Puneet K. Dokania. Multi-agent diverse generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4

Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016. 1, 4

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27 (NeurIPS)*. 2014. 1, 4

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems 30*. 2017. 4

Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 2019. 1, 2

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*. 2017. 2, 4

Quan Hoang, Tu Dinh Nguyen, Trung Le, and Dinh Phung. Mgan: Training generative adversarial nets with multiple generators. In *International Conference on Learning Representations (ICLR)*, 2018. 4

Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles. Learning to decompose and disentangle representations for video prediction. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*. 2018. 1

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2, 4

Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Trans. Graph.*, 2017. 1

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 4

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, 2016. 2

Anwar Kamil and Talal Shaikh. Literature review of generative models for image-to-image translation problems. In *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, 2019. 1

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. 2

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014. 1

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *Proceedings of The 33rd International Conference on Machine Learning (ICML)*, 2016. 4

Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *ECCV*, 2016. 1

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of IEEE conference on computer vision and pattern recognition*, 2017. 1

Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *The European Conference on Computer Vision (ECCV)*, 2018. 2, 4, 6, 8

Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision*, 2020. 6, 7

Jianxin Lin, Yingce Xia, Tao Qin, Zhibo Chen, and Tie-Yan Liu. Conditional image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

Jianxin Lin, Zhibo Chen, Yingce Xia, Sen Liu, Tao Qin, and Jiebo Luo. Exploring explicit domain supervision for latent space disentanglement in unpaired image-to-image translation. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2, 4

Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*. 2016. 4

Yongyi Lu, Yu-Wing Tai, and Chi-Keung Tang. Attribute-guided face generation using conditional cyclegan. In *The European Conference on Computer Vision (ECCV)*, 2018. 1

Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3, 4, 5, 6

Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 4

Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In *2017 International Conference on Learning Representations (ICLR)*, 2016. 1

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 1

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *2018 International Conference on Learning Representations (ICLR)*, 2018. 4

Sanghyeon Na, Seungjoo Yoo, and Jaegul Choo. Miso: Mutual information loss with stochastic style representations for multimodal image-to-image translation. *arXiv preprint arXiv:1902.03938*, 2019. 2

Tu Nguyen, Trung Le, Hung Vu, and Dinh Phung. Dual discriminator generative adversarial nets. In *Advances in Neural Information Processing Systems 30*. 2017. 4

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. 4

Augustus Odena, Jacob Buckman, Catherine Olsson, Tom Brown, Christopher Olah, Colin Raffel, and Ian Goodfellow. Is generator conditioning causally related to GAN performance? In *Proceedings of the 35th International Conference on Machine Learning*, 2018. 4

Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

Eitan Richardson and Yair Weiss. On gans and gmms. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*. 2018. 3

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*. 2016. 4

Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. Pixelcnn++: A pixelcnn implementation with discretized logistic mixture likelihood and other modifications. In *International Conference on Learning Representations (ICLR)*, 2017. 1

Akash Srivastava, Lazar Valkov, Chris Russell, Michael U. Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. In *Advances in Neural Information Processing Systems 30*. 2017. 4

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, koray kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems (NeurIPS) 29*. 2016a. 1

Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *Proceedings of The 33rd International Conference on Machine Learning ((ICML))*, 2016b. 1

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems 30*. 2017. 1

Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. In *2017 International Conference on Learning Representations (ICLR)*, 2017. 1

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*. 2018a. 1

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018b. 1

Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Bryan Catanzaro, and Jan Kautz. Few-shot video-to-video synthesis. In *Advances in Neural Information Processing Systems 32*. 2019. 1

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018c. 1

Po-Wei Wu, Yu-Jing Lin, Che-Han Chang, Edward Y. Chang, and Shih-Wei Liao. Relgan: Multi-domain image-to-image translation via relative attributes. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 2

Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. Diversity-sensitive conditional generative adversarial networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. 2, 3, 4, 5

Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 1

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 1

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017a. 1, 2, 6

Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*. 2017b. 2, 4